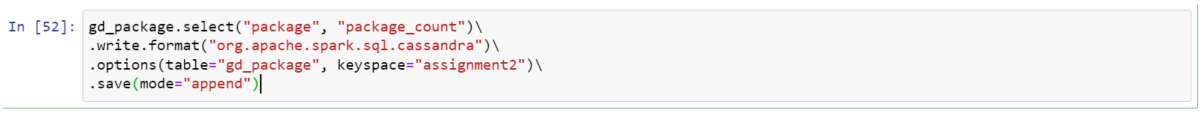
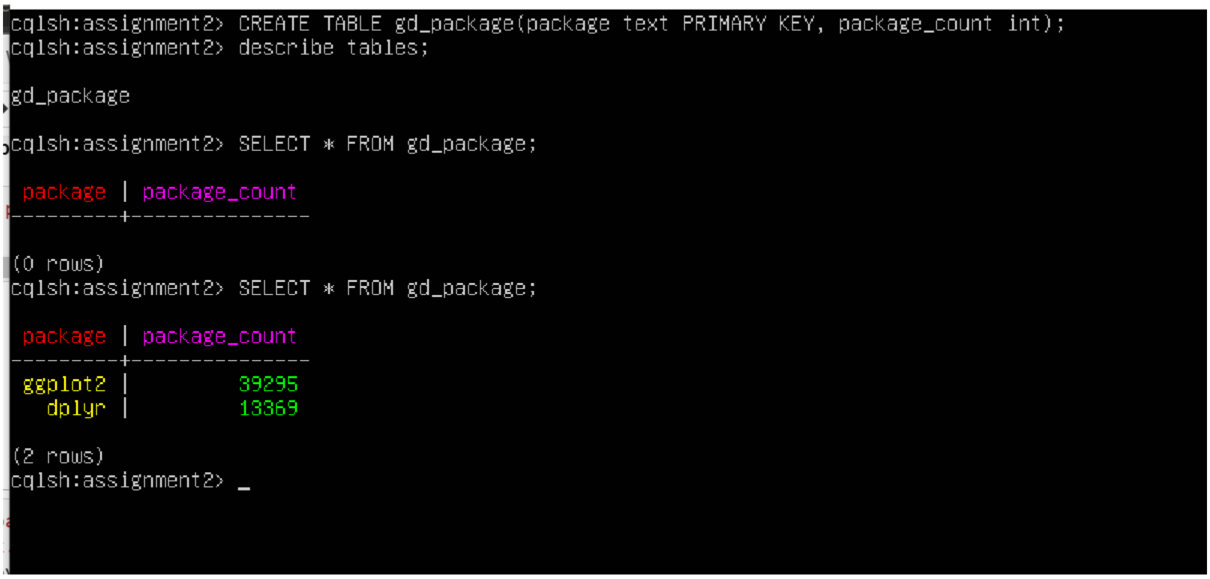
Victor Abiodun Adepoju

3023507

Question 1: Show total number of downloads for packages ggplot2 and dplyr

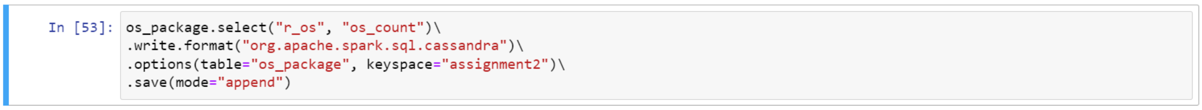


```
In [52]: gd_package.select("package", "package_count")\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="gd_package", keyspace="assignment2")\
         .save(mode="append")
```

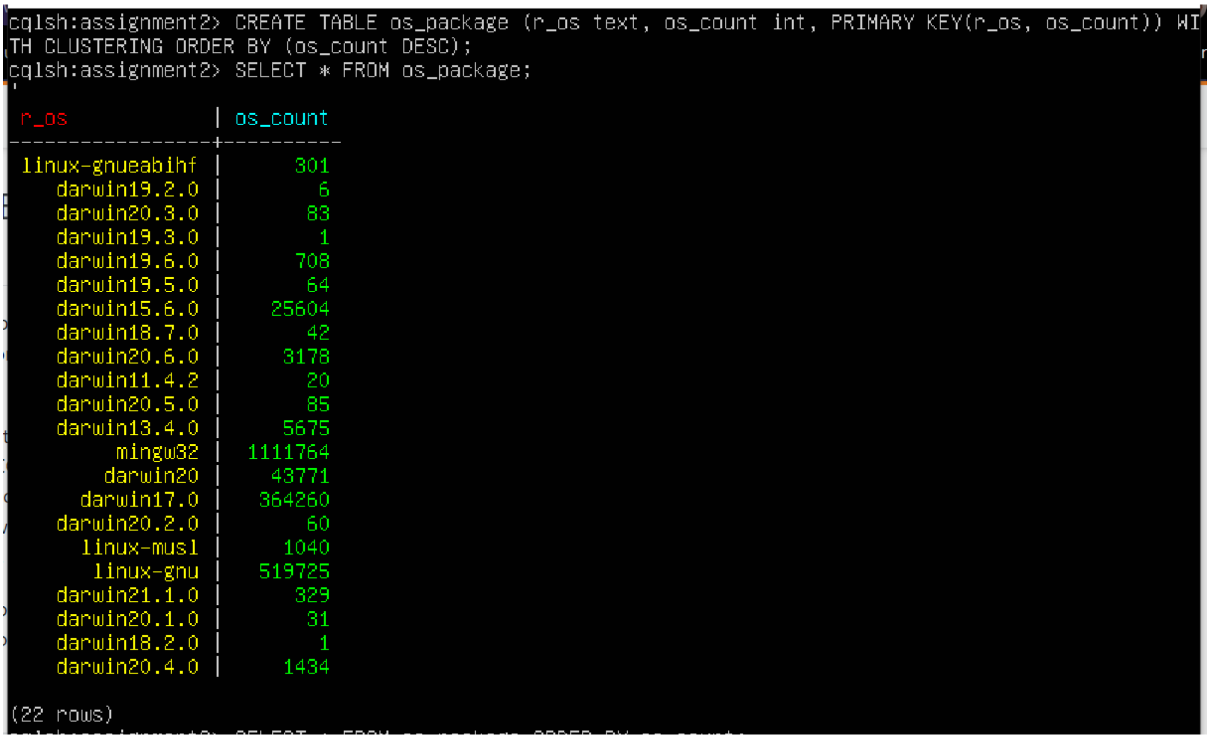


```
cqlsh:assignment2> CREATE TABLE gd_package(package text PRIMARY KEY, package_count int);
cqlsh:assignment2> describe tables;

gd_package

cqlsh:assignment2> SELECT * FROM gd_package;

 package | package_count
---------+---------------

(0 rows)
cqlsh:assignment2> SELECT * FROM gd_package;

 package | package_count
---------+---------------
 ggplot2 |         39295
   dplyr |         13369

(2 rows)
cqlsh:assignment2> _
```

Question 2: Total number of downloads by each Operating System (group similar ones).

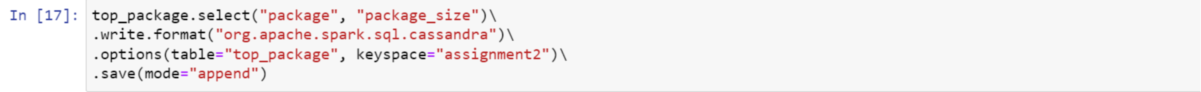


```
In [53]: os_package.select("r_os", "os_count")\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="os_package", keyspace="assignment2")\
         .save(mode="append")
```

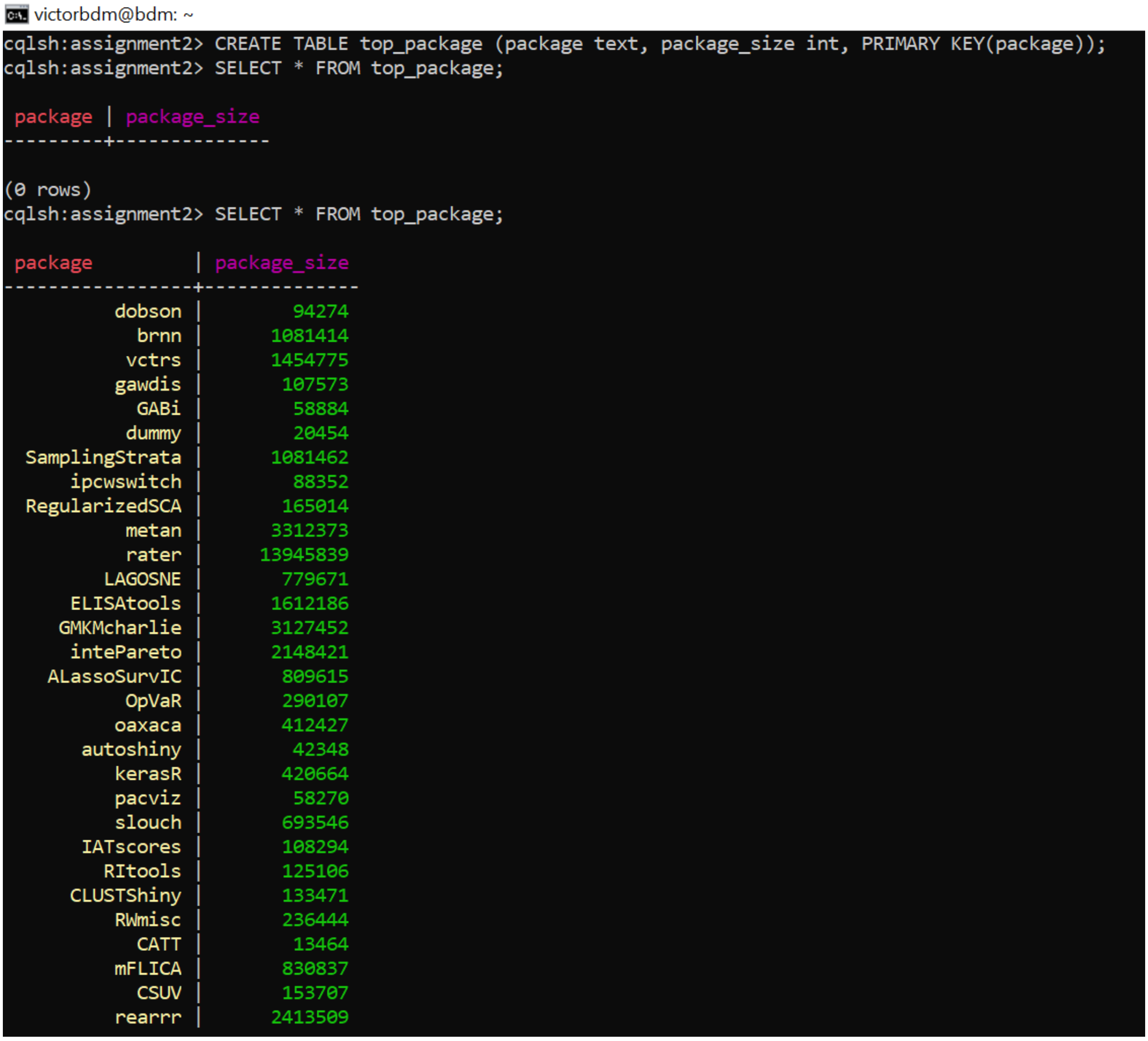


```
cqlsh:assignment2> CREATE TABLE os_package (r_os text, os_count int, PRIMARY KEY(r_os, os_count)) WI
TH CLUSTERING ORDER BY (os_count DESC);
cqlsh:assignment2> SELECT * FROM os_package;

 r_os              | os_count
-------------------+----------
 linux-gnueabihf   |      301
    darwin19.2.0   |        6
    darwin20.3.0   |       83
    darwin19.3.0   |        1
    darwin19.6.0   |      708
    darwin19.5.0   |       64
    darwin15.6.0   |    25604
    darwin18.7.0   |       42
    darwin20.6.0   |     3178
    darwin11.4.2   |       20
    darwin20.5.0   |       85
    darwin13.4.0   |     5675
         mingw32   |  1111764
        darwin20   |    43771
      darwin17.0   |   364260
    darwin20.2.0   |       60
      linux-musl   |     1040
       linux-gnu   |   519725
    darwin21.1.0   |      329
    darwin20.1.0   |       31
    darwin18.2.0   |        1
    darwin20.4.0   |     1434

(22 rows)
cqlsh:assignment2> SELECT * FROM os_package ORDER BY os_count;
```

Question 3: Top 10 (distinct) largest sized packages



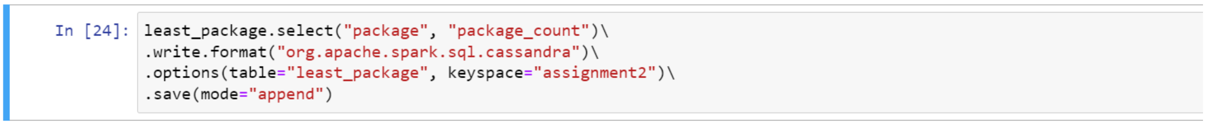
```
In [17]: top_package.select("package", "package_size")\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="top_package", keyspace="assignment2")\
         .save(mode="append")
```



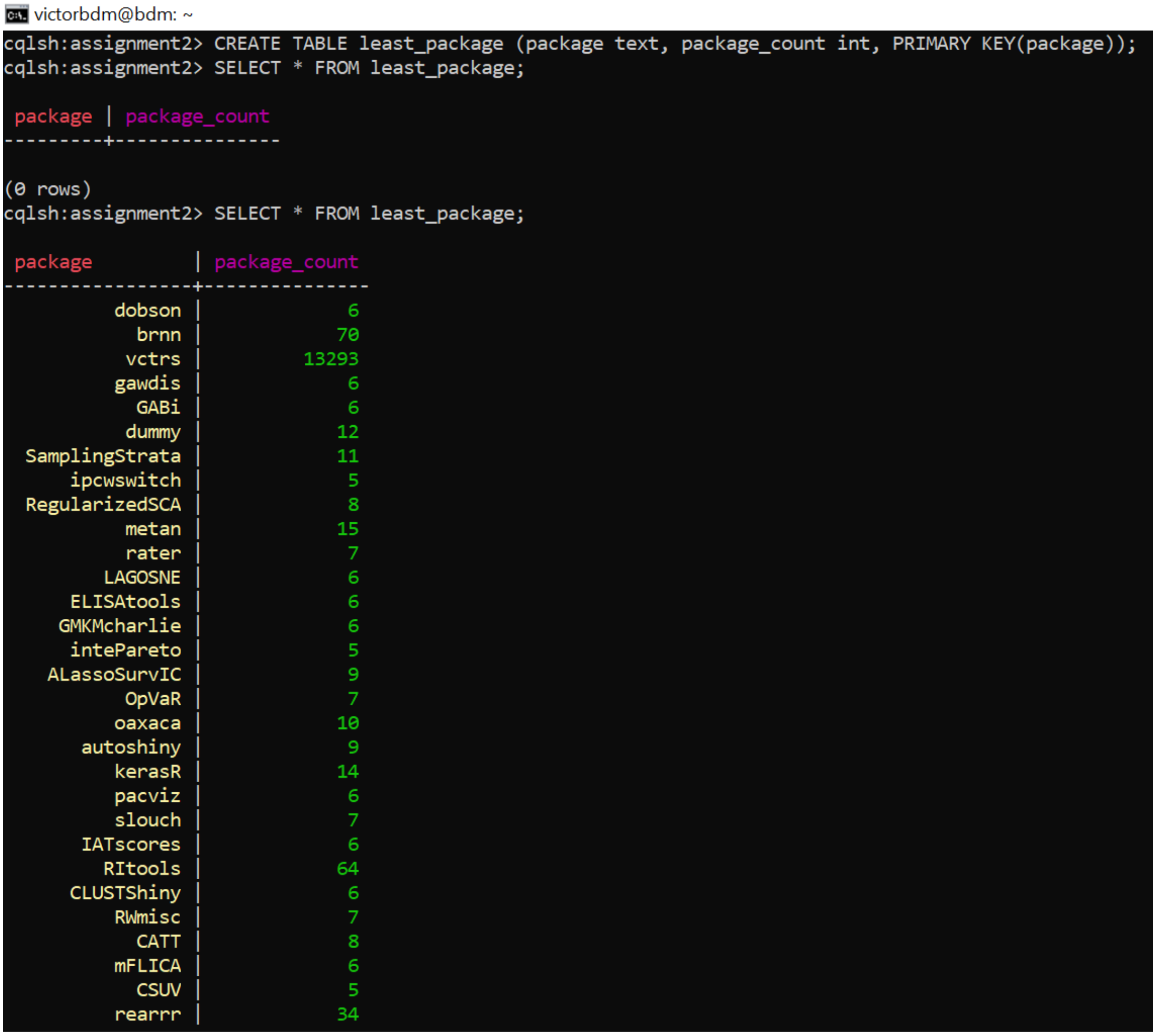
victorbdm@bdm: ~

```
cqlsh:assignment2> CREATE TABLE top_package (package text, package_size int, PRIMARY KEY(package));
cqlsh:assignment2> SELECT * FROM top_package;

 package | package_size
---------+--------------

(0 rows)
cqlsh:assignment2> SELECT * FROM top_package;

 package          | package_size
-----------------+--------------
          dobson |        94274
            brnn |      1081414
           vctrs |      1454775
          gawdis |       107573
            GABi |        58884
           dummy |        20454
  SamplingStrata |      1081462
      ipcwswitch |        88352
  RegularizedSCA |       165014
           metan |      3312373
           rater |     13945839
         LAGOSNE |       779671
      ELISAtools |      1612186
     GMKMcharlie |      3127452
      intePareto |      2148421
    ALassoSurvIC |       809615
           OpVaR |       290107
          oaxaca |       412427
       autoshiny |        42348
          kerasR |       420664
          pacviz |        58270
          slouch |       693546
       IATscores |       108294
         RItools |       125106
      CLUSTShiny |       133471
          RWmisc |       236444
            CATT |        13464
          mFLICA |       830837
            CSUV |       153707
          rearrr |      2413509
```

Question 4: What were the top 10 least popular (distinct) packages?



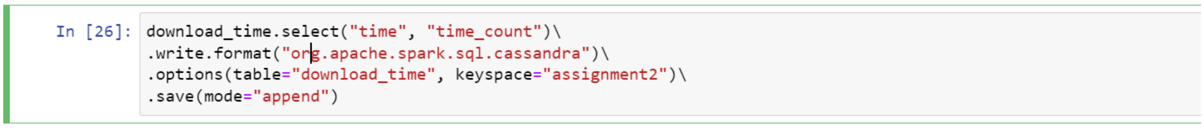
```
In [24]: least_package.select("package", "package_count")\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="least_package", keyspace="assignment2")\
         .save(mode="append")
```



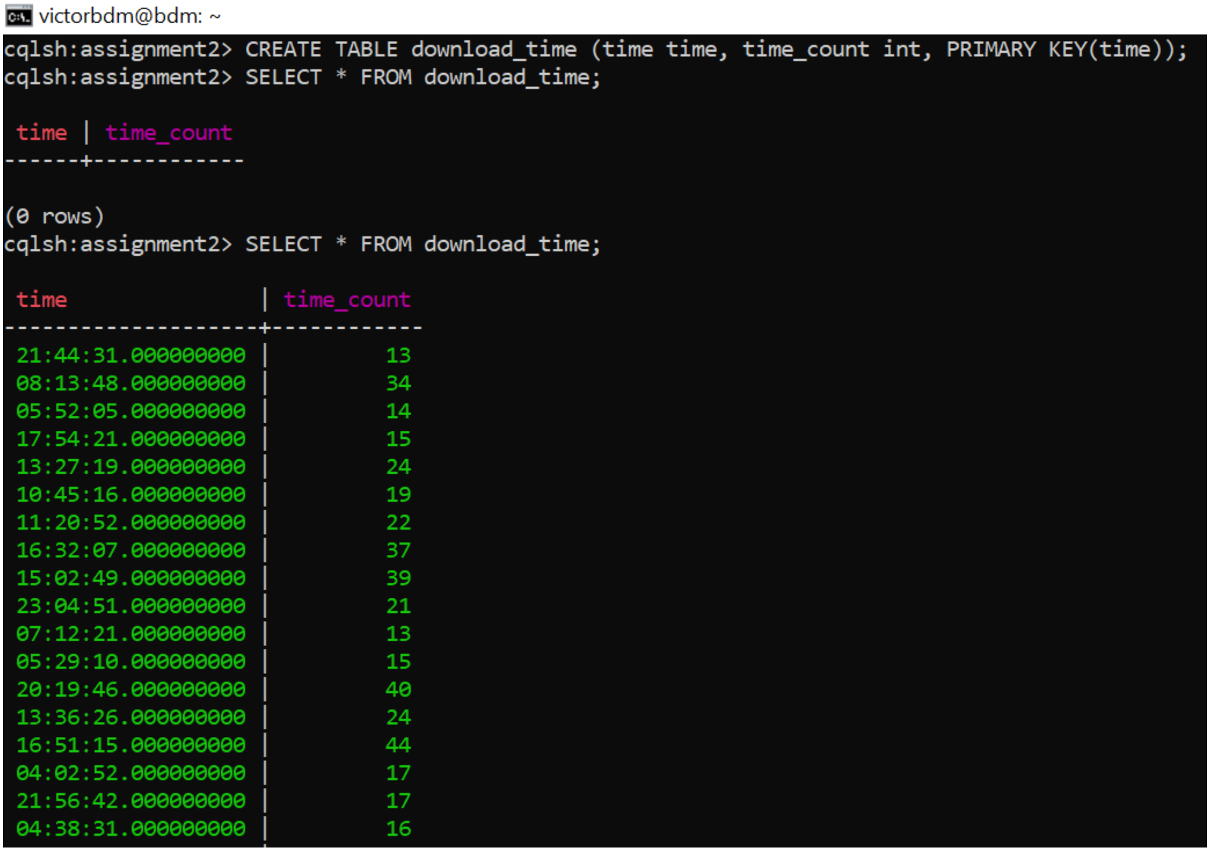
victorbdm@bdm: ~

```
cqlsh:assignment2> CREATE TABLE least_package (package text, package_count int, PRIMARY KEY(package));
cqlsh:assignment2> SELECT * FROM least_package;

 package | package_count
---------+---------------

(0 rows)
cqlsh:assignment2> SELECT * FROM least_package;

 package         | package_count
-----------------+---------------
          dobson |             6
            brnn |            70
           vctrs |         13293
          gawdis |             6
            GABi |             6
           dummy |            12
  SamplingStrata |            11
      ipcwswitch |             5
  RegularizedSCA |             8
           metan |            15
           rater |             7
         LAGOSNE |             6
      ELISAtools |             6
     GMKMcharlie |             6
      intePareto |             5
     ALassoSurvIC |             9
           OpVaR |             7
          oaxaca |            10
       autoshiny |             9
          kerasR |            14
          pacviz |             6
          slouch |             7
       IATscores |             6
         RItools |            64
      CLUSTShiny |             6
          RWmisc |             7
            CATT |             8
          mFLICA |             6
            CSUV |             5
          rearrr |            34
```

Question 5: At what specific hour there are most of the download hits?



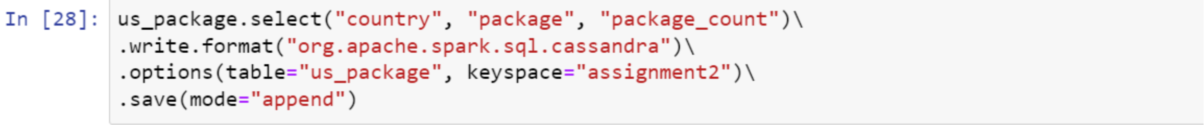
```
In [26]: download_time.select("time", "time_count")\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="download_time", keyspace="assignment2")\
         .save(mode="append")
```



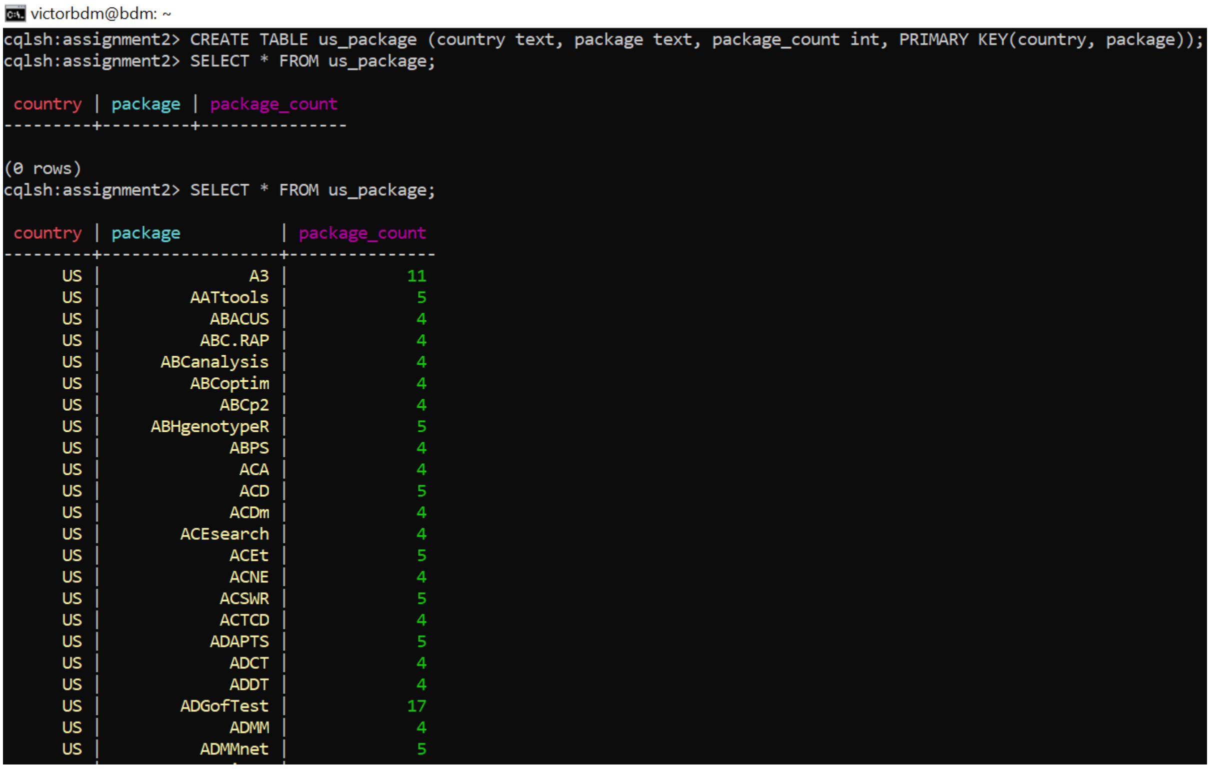
victorbdm@bdm: ~

```
cqlsh:assignment2> CREATE TABLE download_time (time time, time_count int, PRIMARY KEY(time));
cqlsh:assignment2> SELECT * FROM download_time;

 time | time_count
------+------------

(0 rows)
cqlsh:assignment2> SELECT * FROM download_time;

 time                | time_count
--------------------+------------
 21:44:31.000000000 |         13
 08:13:48.000000000 |         34
 05:52:05.000000000 |         14
 17:54:21.000000000 |         15
 13:27:19.000000000 |         24
 10:45:16.000000000 |         19
 11:20:52.000000000 |         22
 16:32:07.000000000 |         37
 15:02:49.000000000 |         39
 23:04:51.000000000 |         21
 07:12:21.000000000 |         13
 05:29:10.000000000 |         15
 20:19:46.000000000 |         40
 13:36:26.000000000 |         24
 16:51:15.000000000 |         44
 04:02:52.000000000 |         17
 21:56:42.000000000 |         17
 04:38:31.000000000 |         16
```

Question 6: . What are the 5 most popular packages in US?



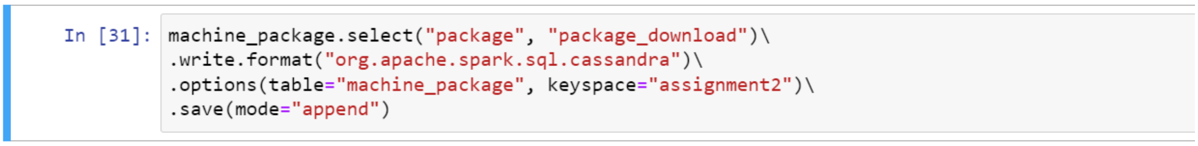
```
In [28]: us_package.select("country", "package", "package_count")\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="us_package", keyspace="assignment2")\
         .save(mode="append")
```



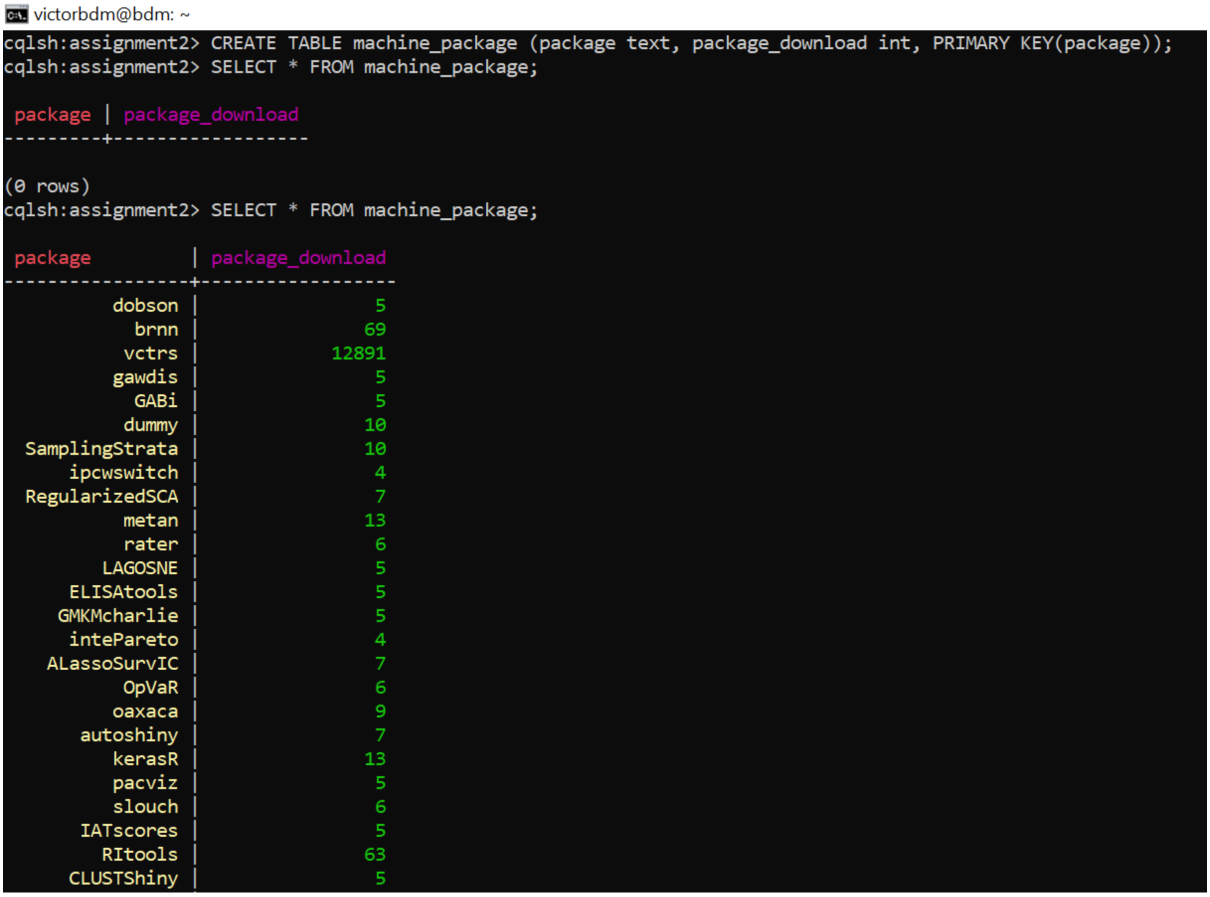
victorbdm@bdm: ~

```
cqlsh:assignment2> CREATE TABLE us_package (country text, package text, package_count int, PRIMARY KEY(country, package));
cqlsh:assignment2> SELECT * FROM us_package;

 country | package | package_count
---------+---------+---------------

(0 rows)
cqlsh:assignment2> SELECT * FROM us_package;

 country | package           | package_count
---------+-------------------+---------------
      US |                A3 |            11
      US |          AATtools |             5
      US |            ABACUS |             4
      US |           ABC.RAP |             4
      US |       ABCanalysis |             4
      US |          ABCoptim |             4
      US |            ABCp2  |             4
      US |       ABHgenotypeR |            5
      US |              ABPS |             4
      US |               ACA |             4
      US |               ACD |             5
      US |              ACDm |             4
      US |         ACEsearch |             4
      US |              ACEt |             5
      US |              ACNE |             4
      US |             ACSWR |             5
      US |             ACTCD |             4
      US |            ADAPTS |             5
      US |              ADCT |             4
      US |              ADDT |             4
      US |          ADGofTest |           17
      US |              ADMM |             4
      US |           ADMMnet |             5
```

Question 7: . Show all packages downloaded by the machine with highest number of downloads?



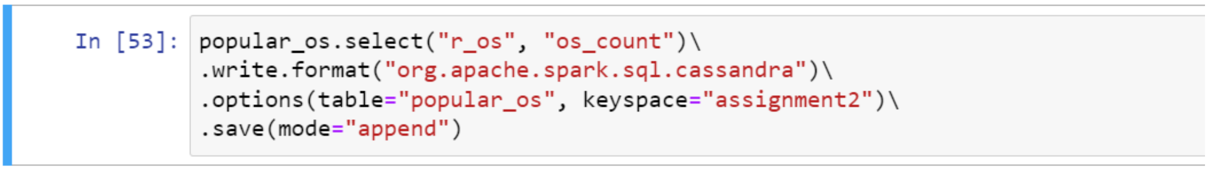
```
In [31]: machine_package.select("package", "package_download")\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="machine_package", keyspace="assignment2")\
         .save(mode="append")
```



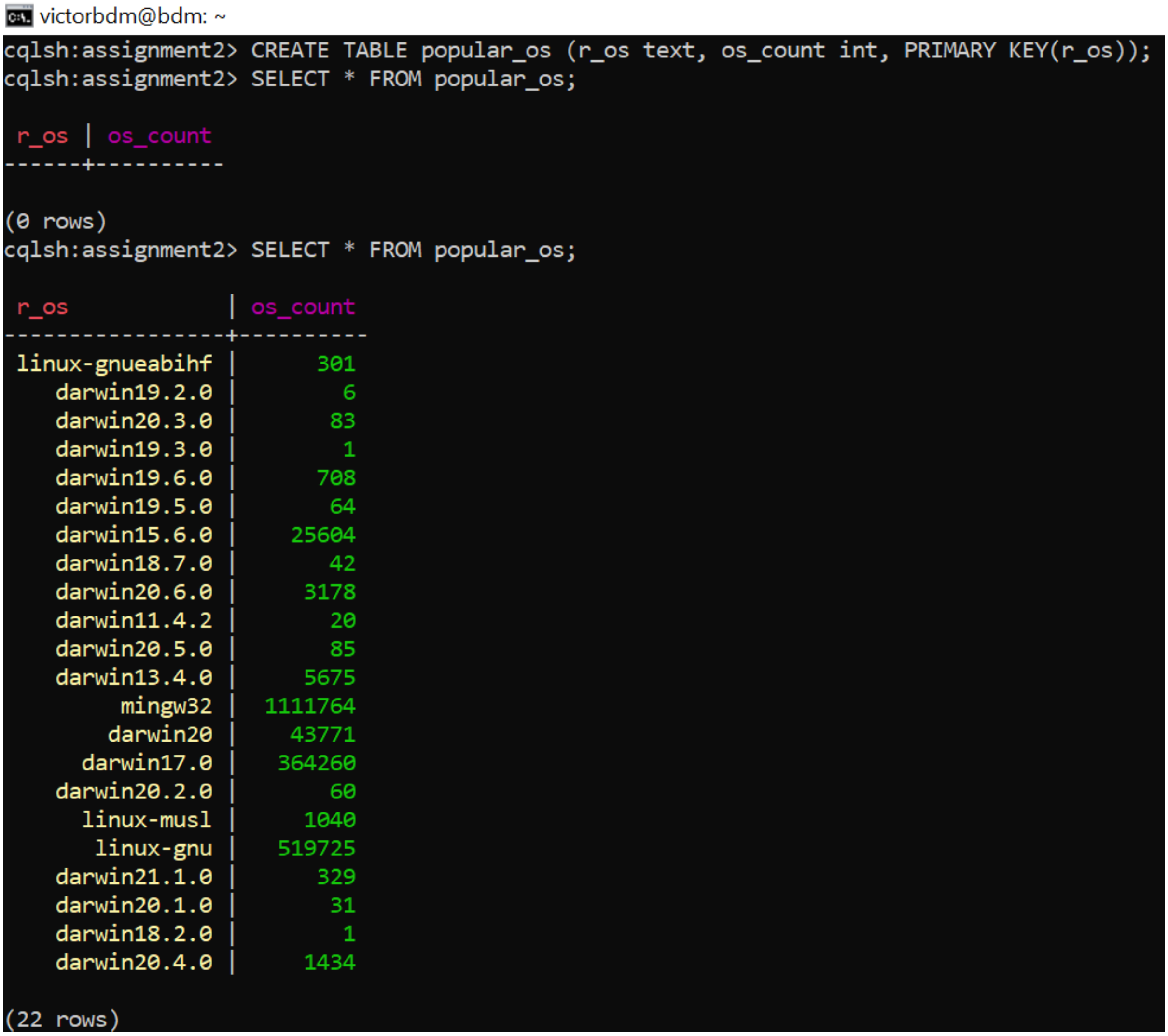
victorbdm@bdm: ~

```
cqlsh:assignment2> CREATE TABLE machine_package (package text, package_download int, PRIMARY KEY(package));
cqlsh:assignment2> SELECT * FROM machine_package;

 package | package_download
---------+------------------

(0 rows)
cqlsh:assignment2> SELECT * FROM machine_package;

 package          | package_download
------------------+------------------
           dobson |                5
             brnn |               69
            vctrs |            12891
           gawdis |                5
             GABi |                5
            dummy |               10
   SamplingStrata |               10
       ipcwswitch |                4
   RegularizedSCA |                7
            metan |               13
            rater |                6
          LAGOSNE |                5
       ELISAtools |                5
      GMKMcharlie |                5
       intePareto |                4
     ALassoSurvIC |                7
            OpVaR |                6
           oaxaca |                9
        autoshiny |                7
           kerasR |               13
           pacviz |                5
           slouch |                6
        IATscores |                5
          RItools |               63
       CLUSTShiny |                5
```

Question 8: . Show top three OSs that are most popular among the R programmers?



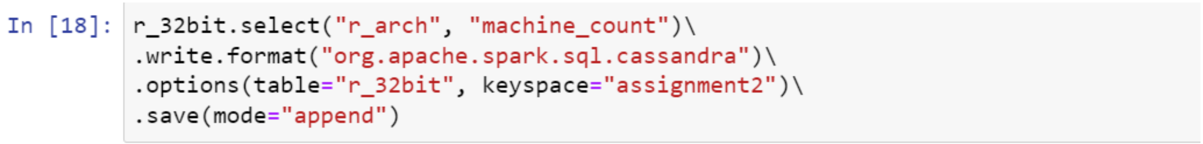
```
In [53]: popular_os.select("r_os", "os_count")\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="popular_os", keyspace="assignment2")\
         .save(mode="append")
```



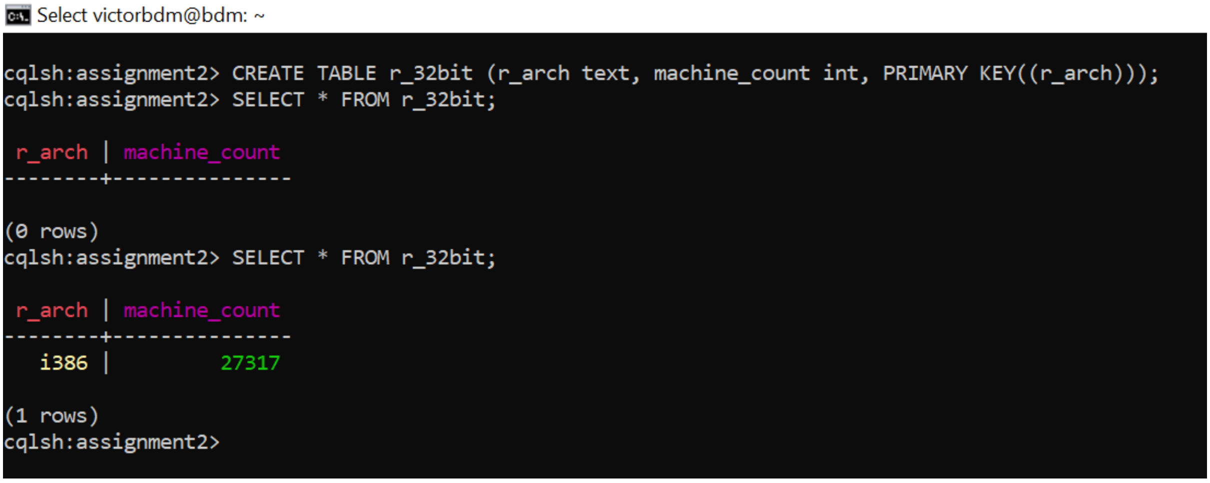
victorbdm@bdm: ~

```
cqlsh:assignment2> CREATE TABLE popular_os (r_os text, os_count int, PRIMARY KEY(r_os));
cqlsh:assignment2> SELECT * FROM popular_os;

 r_os | os_count
------+----------

(0 rows)
cqlsh:assignment2> SELECT * FROM popular_os;

 r_os             | os_count
------------------+----------
 linux-gnueabihf |      301
    darwin19.2.0 |        6
    darwin20.3.0 |       83
    darwin19.3.0 |        1
    darwin19.6.0 |      708
    darwin19.5.0 |       64
    darwin15.6.0 |    25604
    darwin18.7.0 |       42
    darwin20.6.0 |     3178
    darwin11.4.2 |       20
    darwin20.5.0 |       85
    darwin13.4.0 |     5675
         mingw32 |  1111764
        darwin20 |    43771
      darwin17.0 |   364260
    darwin20.2.0 |       60
      linux-musl |     1040
       linux-gnu |   519725
    darwin21.1.0 |      329
    darwin20.1.0 |       31
    darwin18.2.0 |        1
    darwin20.4.0 |     1434

(22 rows)
```

Question 9: . How many R users still use 32 bit machines?

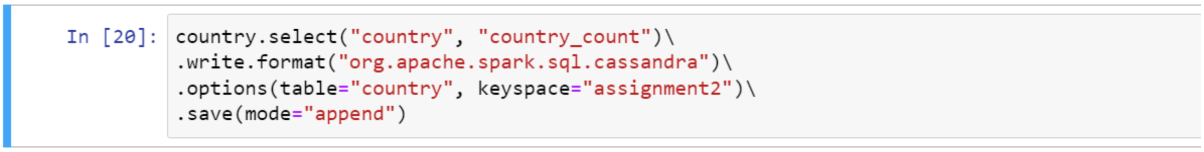


```
In [18]: r_32bit.select("r_arch", "machine_count")\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="r_32bit", keyspace="assignment2")\
         .save(mode="append")
```

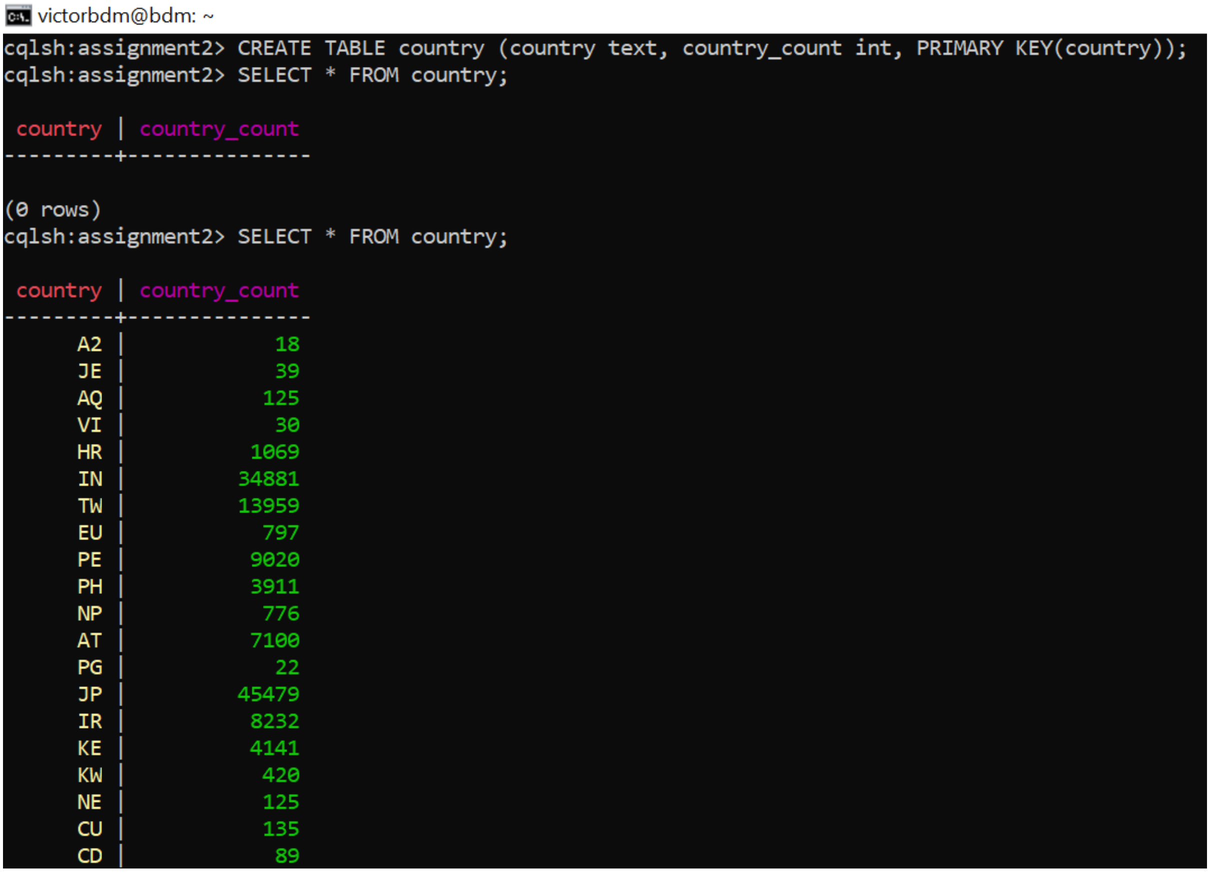


```
cqlsh:assignment2> CREATE TABLE r_32bit (r_arch text, machine_count int, PRIMARY KEY((r_arch)));
cqlsh:assignment2> SELECT * FROM r_32bit;

 r_arch | machine_count
--------+---------------

(0 rows)
cqlsh:assignment2> SELECT * FROM r_32bit;

 r_arch | machine_count
--------+---------------
   i386 |         27317

(1 rows)
cqlsh:assignment2>
```

Question 10: Show total number of downloads by each country, use ascending order?



```
In [20]: country.select("country", "country_count")\
         .write.format("org.apache.spark.sql.cassandra")\
         .options(table="country", keyspace="assignment2")\
         .save(mode="append")
```



```
cqlsh:assignment2> CREATE TABLE country (country text, country_count int, PRIMARY KEY(country));
cqlsh:assignment2> SELECT * FROM country;

 country | country_count
---------+---------------

(0 rows)
cqlsh:assignment2> SELECT * FROM country;

 country | country_count
---------+---------------
      A2 |            18
      JE |            39
      AQ |           125
      VI |            30
      HR |          1069
      IN |         34881
      TW |         13959
      EU |           797
      PE |          9020
      PH |          3911
      NP |           776
      AT |          7100
      PG |            22
      JP |         45479
      IR |          8232
      KE |          4141
      KW |           420
      NE |           125
      CU |           135
      CD |            89
```